AN INVESTIGATION INTO THE USE OF LINGUISTIC CONTEXT IN AUTOMATIC CURSIVE SCRIPT RECOGNITION

Brammall, N.H.; Conolly, J.H.; Hinde, C.J.

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2004

AN INVESTIGATION INTO THE USE OF LINGUISTIC CONTEXT IN AUTOMATIC CURSIVE SCRIPT RECOGNITION

N.H. BRAMMALL, J.H. CONNOLLY AND C.J. HINDE

Department of Computer Studies, Loughborough University, UK.

The highly ambiguous nature of cursive writing, with high variability not only between different writers but also between different samples from the same writer, means that automatic recognition systems based on purely visual information are prone to errors.
It is suggested that the application of linguistic knowledge to the recognition task may improve recognition accuracy. There are many forms of linguistic knowledge that may be used to this end. This paper looks specifically at the use of collocation as a source of linguistic knowledge. Collocation describes the statistical tendency of certain words to co-occur in a language, within a defined range.
The construction and use of a post-processing system incorporating collocational knowledge is described, as are a number of experiments designed to test the effectiveness of collocation as an aid to text recognition.

1 Introduction

Many systems attempting automatic handwriting recognition have based their attempts on purely visual information. While increasingly sophisticated methods have brought about a marked improvement in accuracy, there seems to have been a law of diminishing returns at work. In other words there appears to be a ceiling of accuracy above which methods employing purely visual information cannot go when attempting to recognise handwriting from a wide variety of sources.

The benchmark against which the performance of text recognition systems has traditionally been judged is human performance. Humans have a remarkable (though far from infallible) ability to read even the most visually degraded text. Clearly sources of information above and beyond the merely visual are at work here. As many studies have shown (e.g. Ehrlich & Rayner, 1981, Just & Carpenter, 1987), humans use many levels of knowledge to interpret handwriting and indeed any other image. Pragmatic knowledge, or world knowledge, may place a document in a particular context. At the sentence level, semantic and syntactic knowledge may permit the reader to guess an illegible word by considering the words surrounding it. It is clear that incorporating at least some of these knowledge sources into the automatic recognition of text can offer potential improvements in performance.

In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 475-480

2 Collocation

Collocation is the habitual association of a word in a language with other particular words in that language. A collocation may occur between words in adjacent positions or over a wider frame of reference.

Much of the groundwork for the practical study of collocations was laid by J.McH.Sinclair (Sinclair, 1966). Sinclair suggests that by studying the tendencies of items to collocate with each other, we can discover facts about language that cannot be discovered by grammatical analysis. One item is said to collocate with another item if the probability of it occurring in that item's environment is greater than its individual probability of occurrence would suggest. When studying a text, Sinclair suggests that we measure the way in which an item predicts the occurrence of others and, importantly, also the way in which it is predicted by others.

Statistical tests can assess the significance of any discrepancy between the predicted and the actual number of collocations between two words in a text, giving either a positive correlation, a negative correlation, or an absence of collocation.

One way of setting the level of significance which has been adopted in a number of studies, (Berry-Rogghe, 1973; Lancashire, 1987), is the use of the z-score. In statistics the z-score is a way of ascertaining how many standard deviations from the mean a score lies. In terms of word co-occurrence, the mean can be defined as the number of times two words would be expected to co-occur in a text. If the two words have a strong tendency to co-occur then the z-score representing the probability of this co-occurrence will be high, i.e. the probability of co-occurrence will be a number of standard deviations above the expected probability of co-occurrence. For the method of calculating z-scores, see Berry-Rogghe, 1973.)

3 Input to the System

A component of a text recognition system that exploits linguistic knowledge sources can be viewed as a post-processing 'black box' to a recognition stage acting on visual information.

This low-level recogniser will produce as output a set of hypotheses for each entity in the input image. The role of the post-processor is to select the most appropriate hypothesis for each entity according to the knowledge sources at its disposal. The set of hypotheses for each word position in the sentence is often represented as a word lattice. The post-processing component must choose the path through the lattice which best matches the input image.

Sample word lattices from an existing low-level recogniser were used as the basis for a letter substitution database which could be used in the generation of word lattices.

310 sample word lattices were available giving 34,100 letter substitutions. These substitutions were used as the basis of a word lattice simulation program (see Fig. 1).

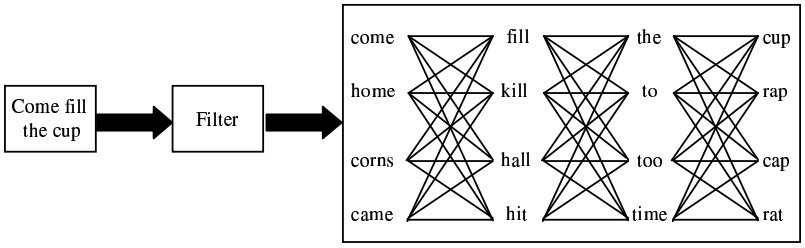

Fig. 1 - A filter to produce a word lattice from a sentence

4 System Components

A word lattice presented to the system is processed with reference to two data sources – a lexicon and a collocation dictionary. The lexicon contains 78,055 words, taken from the Collins Electronic Dictionary. Word look-up is based on the first two letters of the word, allowing words with invalid beginnings to be rejected immediately. The collocation dictionary was compiled by analysing the relationships between words in a section of the British National Corpus, containing around 13 million words (Burnard, 1995).
The collocational relationships between a group of words in a text are illustrated in Fig. 2.

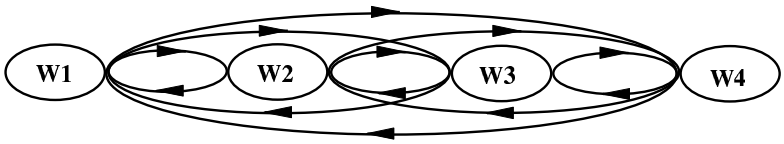

Fig. 2 – Representation of the collocational relationships between four contiguous words.

The two-way nature of a collocational relationship is captured by use of a cascade structure based on alphabetical order. Given two collocationally linked words, the first word alphabetically is considered as an 'anchor', to which the other word's position is relative. Each relationship between words is represented by a single entry capturing the two-way nature of the relationship.

5 The Processing of a Word Lattice

After pre-processing we have a word lattice consisting of the input sentence plus four valid candidate words for each word position in that sentence. Initially, the first five word positions in the lattice are considered, giving 3125 (5^5) potential paths through the lattice. Each of these is analysed in sequential order, using a pipeline mechanism. For the path being analysed, the first word alphabetically is identified, and treated as the node. The collocation dictionary is searched for collocations between the node and each of the other words in the sequence. If a match is found then the collocational score between the node word and that collocate is calculated. The strength of the collocation is represented by two values, due to the two-way nature of collocation. The collocational score of a relationship is calculated by multiplying these two values together. This process is repeated for each word in the path until all the scores have been calculated and added together, giving a score for the path. The scores for each path are compared to find the highest. The first word of the path with the highest score is assigned as the system's hypothesis for the correct word in the first word position.

The next word can now be fed into the pipeline, creating another set of paths. The collocation scores from the previous calculation are retained, as the collocational influence of the word in position number one on the next four word positions must be considered in conjunction with the relationships between these words and the new word in the pipeline. The process is repeated, with the collocation score for each path being added to by the influence of the new word in the pipeline. Once all the scores have been calculated, the path with the highest score provides us with the system's hypothesis for word position two. This continues until the last word position in the lattice has been reached, and therefore a hypothesis has been proposed for each word position in the lattice. These hypotheses are compared with the initial input sentence, to give a measure of success.

6 Experiments

A number of experiments were carried out to test the system using various different criteria. Two variables were considered - the input text and the representation of the strength of a collocation. For input texts, the BNC (used to compile the collocation statistics) and the Susanne Corpus (Sampson, 1995) were used. This effectively presented the system with trained and untrained input. In one set of experiments, collocational strength was represented by z-scores, with only significant collocations being stored. In another set of experiments all collocations in the training text were represented by the percentage of occurrences of a node in collocation with a word in relation to the total number of occurrences of the node in the text.

Fifty sentences containing a total of 338 words were randomly selected from the same subset of the BNC that was used to create the collocation knowledge base. Fifty sentences containing a total of 404 words were randomly selected from the Susanne Corpus. The rates of successful recognition are shown in Fig. 3.

		Input Text	
		BNC	Susanne Corpus
Representation of Collocational Strength	% - score	94.08%	70.79%
	z - score	79.71%	60.05%

Fig. 3 - Percentage of correct word hypotheses given input from two sources using two representations of collocation.

7 Summary

Clearly the performance of the system when working on sentences from the BNC is considerably better than on sentences from the Susanne corpus. Testing the system with sentences from the BNC is equivalent to training the system, as the collocation knowledge base was constructed using statistics extracted from the BNC. Giving the system sentences from the Susanne corpus is equivalent to using the system 'blind', i.e. with no prior knowledge of the input. The percentage of words correctly hypothesised when working with sentences from the BNC suggest that a trained collocational knowledge base could be highly effective in handwriting recognition. The collocation dictionary which uses percentage scores to represent collocations clearly out-performs the z-score dictionary. This is because of the complete coverage of the percentage score dictionary - *all* collocations in the training text are represented. In the z-score dictionary, only those collocations above a specified significance threshold are represented. This means that rarely occurring but valid collocations may slip through the net. It should be noted that processing with the percentage score dictionary was very much slower than with the z-score dictionary, so we have a trade-off between operational accuracy and processing speed.

8 Conclusions

The results obtained strongly suggest that the use of a collocation knowledge base in a post-processing capacity can indeed enhance the performance of a handwriting recognition system. Inevitably mistakes are made, and the reasons behind these errors provide an insight into the nature of collocation.

Of most interest are the errors which occur due to the difference in the nature of collocation when it involves 'grammar' words (frequently-occurring words with a largely grammatical role) as opposed to 'lexical' words, or a combination of the two. Based on the results these experiments, the collocational relationships between lexical words tend to be quite stable and reliable indicators as to the collocational patterns prevalent in a text. Considering the relationships between grammar words and lexical words can, however, give a rather distorted view of these patterns of co-occurrence, due to the differing relative frequencies of grammar words and lexical words. Many of the errors made by the system arose from hypothesising a frequently-occurring grammar word in place of a less frequently-occurring lexical word. Performance may be improved by considering some sort of weighting system based on the frequency of a word in a text.

References

Berry-Rogghe, G.L.M. 1973. 'The Computation of Collocations and their relevance in Lexical Studies', in Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (eds.), pp.103-112.

Burnard, L. (ed.) 1995. *Users Reference Guide for the British National Corpus*. Oxford : Oxford University Computing Services.

Ehrlich, S.F. Rayner, K. 1981. 'Contextual Effects on Word Perception and Eye Movements during Reading', *Journal of Verbal Learning and Verbal Behavior*, 20, pp.641-655.

Hull, J.J. 1994. 'Language-Level Syntactic and Semantic Constraints Applied to Visual Word Recognition', *NATO ASI Series F, Computer and System Sciences*, 124, pp.289-312.

Jones, S. and Sinclair, J.McH. 'English Lexical Collocations', *Cahiers de Lexicologie*, 24, pp.15-61.

Just, M.A. and **Carpenter, P.A.** 1987. *The Psychology of Reading and Language Comprehension*. Newton, Mass. : Allyn and Bacon.

Lancashire, I. 1987. 'Using a Textbase for English-language Research', *Proceedings of the 3rd Annual Conference of the UWC for the New Oxford English Dictionary*, Waterloo, pp.51-64.

Rose, T.G. and **Evett, L.J.** 1995. 'The use of Context in Cursive Script Recognition', *Machine Vision and Applications*, 8, no 4, pp.241-8.

Sampson, G. 1995. *English for the Computer*. London : Oxford University Press.

Sinclair, J.McH. 1966. 'Beginning the Study of Lexis', in Bazell, C.E. *et al*., pp.410-430.